Terabit-scale hybrid networking

SC program title:	Terabit networking for extreme-scale science,
FOA number	DE-FOA-0000523
Applicant/Institution:	University of Virginia
Street Address/City/State/Zip	POB 400195, Charlottesville, VA 22904
Principal Investigator	Malathi Veeraraghavan
Position of PI	Professor
Postal address of PI	POB 400743, Charlottesville, VA 22904
Telephone Number of PI	(434) 982 2208
Fax number of PI	(434) 924 8818
Email address ofPI	mv5g@virginia.edu
DOE OS Program Office	Office of Advanced Scientific Computing Research
Program Office Technical contact:	Dr. Thomas Ndousse-Fetter
Is this a Collaboration	Yes, ESnet (PI: Chris Tracy)

Project objective

The objective of this project is to advance the design of, and to prototype and demonstrate, an effective hybrid network traffic engineering system.

Executive Summary

This application proposes to design, prototype and test multiple versions of a hybrid network traffic engineering system (HNTES) that isolates science flows from general-purpose flows, and configures routers so that the flows are automatically redirected to high-rate circuits. Deployment of HNTES would allow network operators to manage their parts of the end-to-end path more efficiently. Such a system is required because not all operators offer both IP-routed and circuit services, and most applications generate traffic for the IP-routed service. Therefore, in order for a core network operator, such as ESnet, to leverage its circuit services, a system such as HNTES is useful.

Three tasks are executed by HNTES: (i) identification of heavy-hitter flows, (ii) initiation of circuit provisioning, and (iii) configuration of policy based routes (PBR) at ingress and egress IP routers of the operator's network. Of the four dimensions on which flows can be classified: size, rate, duration and burstiness, this work focuses on the size dimension because the DOE scientific flows are primarily large dataset transfers. Three algorithms are proposed for elephant (large-sized) flow identification: (i) offline Netflow data analysis, (i) online flow analysis, and (iii) end-host assisted mechanisms. Three approaches are proposed for circuit provisioning through the IDC: (i) rate-unlimited static MPLS LSPs

initiated offline, (i) rate-unlimited static MPLS LSPs initiated online, and (iii) rate-specified MPLS LSPs initiated online. For the third task, the PBRs can be configured offline or online.

The University of Virginia is **currently funded** by DOE ASCR to develop HNTES. Completed work and ongoing work are clearly described in this application. The first version, *HNTES 1.0*, which was designed, prototyped, tested on the ANI 100G testbed, and demonstrated in Oct. 2010, handled long-duration flows and created dynamic circuits. The following issues were identified with the HNTES 1.0 solution: (i) As heavy hitters, science flows dominate in the size dimension and not the duration dimension (i.e., consuming an unfair share of network resources); (ii) The solution requires HNTES to keep pace with high-rate data-plane packets mirrored to it from routers to run its online heavy-hitter flow detection algorithms, which is impractical without high-performance computing platforms; and (iii) Circuits need to be static (pre-provisioned) because the durations of large-sized flows are typically smaller than the circuit provisioning delay, which is on the order of minutes using the OSCARS IDC. Therefore, the current version, *HNTES 2.0*, identifies elephant flows (size dimension instead of duration dimension) and uses static circuits (instead of dynamic as in HNTES 1.0). A *novel idea* of rate-unlimited static LSPs is used to maintain high throughput for scientific data transfers. An offline elephant flow detection algorithm has been developed and coded, and is being tested on ESnet Netflow data. Data analysis of Netflow data from multiple months will be carried out to evaluate the efficacy of this offline-only approach. Further, three techniques are being used to study the question of why elephant flows should be moved off the IP-routed network: experiments across ESnet, simulations, and analysis of PerfSONAR OWAMP and Netflow data. Preliminary results are presented in this application. These analyses, simulation and experimental studies will be completed, and results will be published in leading conferences and journals. The HNTES 2.0 design will be completed, tested on the ANI 100G testbed, and demonstrated. All these activities are part of the currently funded project.

The **new work** proposed in this application consists of the following: (i) Design of new online flow detection algorithms, (ii) HNTES 3.0 design, prototyping and testing (superset of HNTES 2.0), (iii) Design of end-host assisted flow identification mechanisms, (iv) HNTES 4.0 design, prototyping and testing (superset of HNTES 3.0), (v) HNTES in an integrated network, which requires the design of optimal algorithms for load balancing science flows and sizing LSPs, and (vi) Study of other types of heavy-hitter flows. The offline-only and static LSPs solution of HNTES 2.0 will be augmented with online flow detection algorithms conducted within a provider network (HNTES 3.0), and with assistance from end host flow identification modules (HNTES 4.0). How HNTES can be used in an integrated network, as in one of the options being considered for the next-generation 100 Gbps ESnet, is addressed next. This requires the design of optimal algorithms for load balancing science flows across the whole network topology, and for sizing LSPs. Finally, heavy-hitter flows on the other three dimensions: rate, duration and burstiness, will be considered in studies that evaluate their presence in ESnet, and their effect on general-purpose flows.

Our **evaluation plan** with division of tasks between the two participating institutions, University of Virginia and ESnet, and milestones/deliverables, are presented with a timeline. The project management plan is to continue our current approach in which the two institution PIs and graduate students meet weekly via skype video calls.

Table of Contents

1 Project scope

If proposals can be classified into three types: (i) fundamental research, (ii) applied research/engineering, and (iii) advanced deployment, this proposal falls in the Applied Research/Engineering category.

Of the challenges listed in the solicitation [1], this proposal addresses the following:

- isolate high-impact science flows from normal traffic flows
- allow flows to seamlessly move between shared Internet Protocol (IP) and dynamic circuit infrastructures
- allow network operators to effectively manage their part of the end-to-end path

Of the technical areas listed in the solicitation [1], this proposal addresses the Federated terabit network services and tools topic. Within this topic, of the three current priorities listed in [1], it falls under the scope of the priority:

- Advanced multi-layer and multi-domain services and tools to enable dynamic hybrid networking capabilities based on the emerging 100 Gbps link technologies

2 Background

ESnet services and networks: ESnet currently offers its customers both IP-routed services and circuit services[a]. ESnet has deployed the On-Demand Secure Circuits and Advance Reservation System (OSCARS) Inter-Domain Controller (IDC) to support its circuit services [2]. The IDC supports both a human interface and a programmatic interface to allow users and applications to reserve, provision and release circuits as needed. The reservation requests specify the rate, duration and endpoints of the circuit.

These two types of services can be offered on *separate* networks, or on one *integrated* network. In the current deployment, ESnet4 operates a network of nodes configured as IP routers for the IP-routed service, and a network of separate nodes configured as MultiProtocol Label Switching (MPLS) Label Switched Routers (LSRs) for the circuit services. Links between the nodes of the two networks use separate wavelengths. For the next-generation network, ESnet plans to upgrade to 100 Gbps links between Points of Presence (PoPs). Since network systems from Juniper and other vendors can be configured to operate in both IP-routed and MPLS modes, it is feasible to design a single integrated network, with one node at each PoP and a single set of links between PoPs.

Hybrid Networking: In the DOE context, a *hybrid network* is one that supports both IP-routed and circuit services. It could consist of two separate networks as in ESnet4, or a single integrated network as explained above. A *hybrid network traffic engineering system (HNTES)* is one that moves data flows between these two services as needed [3], i.e., engineers the traffic to use the service type appropriate to the traffic type. Questions that

[a] Strictly speaking, it is a virtual-circuit service, as connection-oriented packet switches, such as MultiProtocol Label Switching (MPLS) systems are used. For simplicity, we refer to this service type as "circuit services."

follow are "what does it mean to move data flows" and "why is it necessary to move data flows between these two services?"

The *answer to the first question* is that a flow appears at an ingress router of a provider's network as an IP-routed flow because most applications generate data for the IP-routed service. Site networks, i.e., campus networks of national laboratories, for which ESnet offers backbone connectivity typically do not offer circuit services, and hence most applications have not been modified to explicitly choose between these two services. Therefore, to move a flow explicitly to a circuit that spans the backbone network (e.g., ESnet), the ingress IP router of the network needs to have been configured with a policy based route (PBR) that allows it to identify all packets corresponding to a particular flow and move those packets to a particular circuit. Most IP routers support this PBR capability whereby a flow can be identified by some subset of the 5 tuples: {destination IP address, source IP address, IP header protocol field, destination transport-layer port number, and source transport-layer port number}. This answers the "what does it mean to move data flows" question.

The *answer to the second question*, "why is it necessary to move data flows between these two services" is that science flows are "heavy-hitter flows" in that their networking resource needs are significantly greater than those of general-purpose flows, and sometimes routing both science and general-purpose flows on shared IP-routed links can cause the latter to experience degraded performance.

Lan and Heidemann [4] classify flows on four dimensions: size (elephant and mice), rate (cheetah and snail), duration (tortoise and dragonfly), and burstiness (porcupine and stingray). Of these, this work focuses on the size dimension since as noted in [5], leadership-class computing facilities, such as OLCF, are facing a need for exponential growth in storage capacity as shown in Figure 1. For example, currently OLCF has 16 PB, and the growth rate is 30 TB per day [5]. As computing platforms increase in speed from petaflops to exaflops, scientists can execute a larger variety and a greater number of simulation runs. Correspondingly, this leads to an increasing volume of the computed, and hence stored, data. These large datasets will need to be transferred to the scientists' sites from the sites of the computing facilities to allow for local analysis and storage.

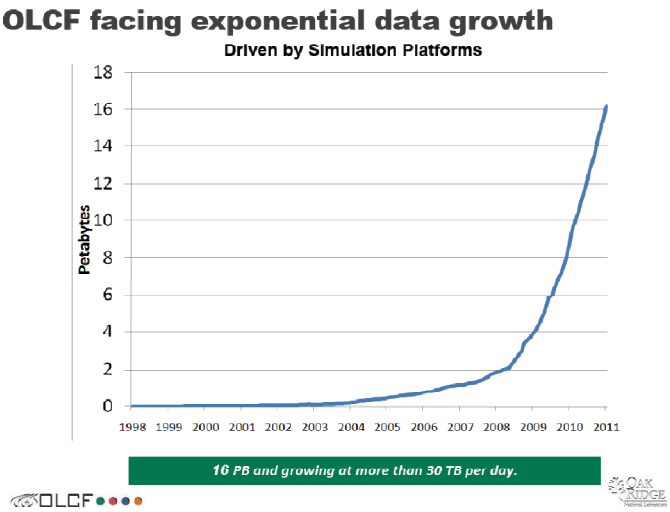

Figure 1: Growth in required storage at Oak Ridge Leadership Computing Facility (OLCF) [5]

Along with simulation-generated datasets, new experimental facilities, such as the National Synchrotron Light Source II (NSLS-II), International Thermonuclear Experimental Reactor - ITER, Large Hadron Collider - LHC, Linac Coherent Light Source - LCLS) [1] will produce large amounts of data. A recent paper on Climate Science notes that there are 4 types of

data: (i) simulation models generated data, (ii) in-situ (instrument) data from sensors, (iii) observational data, and (iv) reanalysis data [6]. In summary, the main dimension of heavy-hitter flows that is of interest to the DOE community is size (see workshop reports [7]) .

Elephant (large-sized) flows impact general-purpose flows significantly because of TCP's congestion control algorithm, which allows the congestion window at the TCP sender to grow until the full link rate (of the bottleneck link on the end-to-end path) is exploited. As mice flows often complete before reaching this limit, the throughput of mice flows is smaller than that of elephant flows. This unfairness is documented in a paper titled, "The war between elephants and mice" [8]. Also, delay-sensitive flows experience increased jitter in the presence of elephant flows whose TCP senders send back-to-back packets causing router buffer occupancy to increase thereby slowing down packets from delay-sensitive flows. This answers the second question as to why flows are moved.

Importance of HNTES in Terabit networks: Hybrid networking will become more important as we transition to Terabit/s networking because file sizes will grow as more simulations are run by the scientific community on the significantly faster exascale computing platforms (relative to today's petascale platforms). Since small messages will still be required for services such as DNS lookups, the variability in message/file sizes will increase significantly in terabit networks. This makes it even more important than it is today to isolate science flows from general-purpose flows.

3 Problem statement and research challenges

Problem statement: ESnet reported recently that 50% of its traffic consists of scientific flows, and that it has a need to "automate portions of the process of identifying candidate large flows and re-routing them over OSCARS layer3 circuits" [9]. As most of the site and peer networks do not offer circuit services, ESnet and other network operators who offer both types of service should be provided the tools necessary to manage their part of the end-to-end paths effectively (as per one of the challenges listed in the solicitation [1]).

The *problem statement* of this proposed project is to design, prototype and evaluate innovative algorithms and systems for hybrid network traffic engineering. The ultimate goal is deployment in the DOE terabit networking infrastructure. Hence this study will make assumptions that accurately reflect realities and constraints of today's networks [10].

Research challenges: The first research challenge is to analyze Netflow data in ESnet and DOE laboratory networks to determine whether offline flow classification is sufficient, or if online classification is required. The former is possible if the identifiers of heavy-hitter flows remain unchanged, and if the rate of arrival of new unexpected heavy-hitter flows is small. If offline classification is sufficient, the next set of research challenges are to develop algorithms and network management systems to size static circuits (determine appropriate rates) and to select values for parameters that define "heavy-hitter" flows (e.g., is a 1 GB data transfer an elephant flow?).

If online classification is required, the problem is more complex. Online flow classification poses system design and implementation challenges. The flow classification, circuit setup

and route configuration steps have to all be executed before the flow terminates, and large-sized flows can be quite short-lived. This makes it a highly challenging systems problem.

4 Related work

Three types of traffic classification methods are described in a review paper [11]: (i) Port based, (ii) Payload based and (iii) Machine Learning techniques. The first set of methods are dismissed for general-purpose IP traffic because P2P applications use port masquerading, and payload analyses do not work with encrypted payloads. Machine learning approaches include Bayesian analysis [12], K-Means [13], Naive Bayes tree [14], Bayesian neural networks [15], and decision trees [16], among others.

But for the DOE scientific community usage, the simpler techniques of packet-header based and payload based analysis can be used. The data transfer nodes at major computing facilities such as NERSC, ALCF, OLCF, and other DOE laboratories, have static well-known public IP addresses. Similarly, file transfer tools such as GridFTP and BBFTP use authentication methods but not encryption of their packets, and so payloads of packets carried on file transfer application control connections can be inspected.

Other projects that address this question of interfacing with circuit services include Lambdastation [17], Terapaths [18], ESCPS [19], and Phoebus [20]. Papers on multi-layer architectures, such as [3], [21], [22], are important to HNTES. For our work on optimal algorithms for static LSP routing, findings from [23], [24], and [25] will be leveraged.

Relationship with other proposals: Another proposal, led by Argonne National Laboratory, in which both University of Virginia and ESnet are participants, addresses the question of how to integrate Globus Online with HNTES. It leverages macroscopic knowledge available to Globus Online to notify HNTES servers about whether or not their networks are bottlenecks for high-throughput performance on end-to-end paths. The focus in that proposal is on directly improving scientific data transfer performance. In contrast, this proposal focuses on the development of the HNTES system to identify and reroute heavy-hitter flows off the IP-routed path as some of the scientific flows have a significant negative impact on general-purpose flows. Therefore, the work planned in these two proposals is strictly complementary.

5 A Hybrid Network Traffic Engineering System (HNTES) architecture

5.1 Role of HNTES

We start with a big picture illustration of what role HNTES would play if deployed by those providers who offer both IP-routed and circuit services. In the example shown in Figure 2, the "DOE Lab I" network and ESnet are shown as offering both IP-routed and circuit services, while the "DOE Lab II," "Other provider" and "University" networks only offer IP-routed services. This reflects current-day reality. An IDC, which offers circuit scheduling, provisioning and release functionality, and a HNTES system are shown as being deployed in DOE Lab I and ESnet.

If a large dataset transfer is initiated from the data transfer nodes at DOE Lab II to the data transfer nodes at DOE Lab I, then the elephant flows will appear as IP-routed packets at the ESnet router R3. Without HNTES, packets from these flows would be forwarded with the IP-routed service via ESnet and DOE Lab I's network. With circuit service, there are two possibilities, intra-domain circuits and inter-domain circuits. Here, we focus on the former, though as the IDCP (IDC Protocol) [26] becomes more commonly deployed, HNTES can help coordinate the use of inter-domain circuits as well.

If offline flow analysis had identified this particular flow (e.g., if the source IP address and destination IP address had occurred often in prior Netflow records), then an MPLS LSP would have already been configured between router R3 and router R2 passing through LSRs S3, S2 and other intermediate LSRs for the intra-domain circuit through ESnet, and PBRs would have already been configured at routers R3 and R2 to forward packets from this flow to this circuit. If the flow was unexpected (i.e., it had not been seen previously) a live identification of this flow as a potential elephant flow, a live circuit setup and a live PBR configuration of the routers are required. These tasks are elaborated upon in the next paragraph.

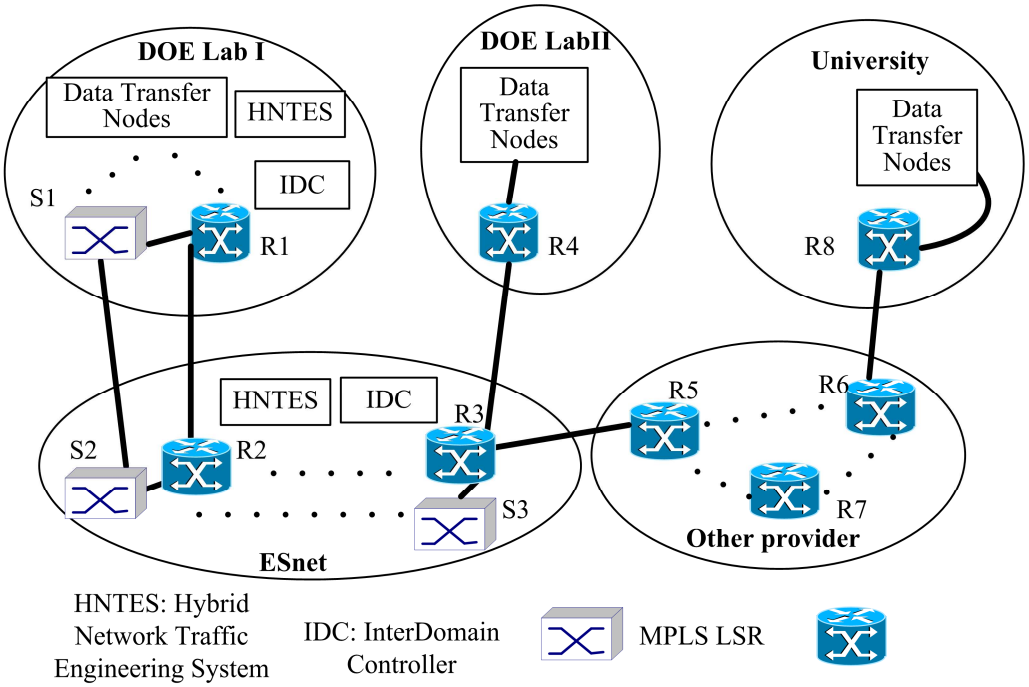

Figure 2: Role of HNTES: An example deployment scenario

5.2 Tasks executed by HNTES

As shown in Figure 3, three types of tasks are executed by HNTES: elephant flow identification, circuit provisioning, and policy based route (PBR) configuration. Three approaches are identified for *elephant flow identification*: offline flow analysis, online flow analysis, and end-host assisted. In offline flow analysis, Netflow data collected by the IP routers (which is sampled packet header data) is analyzed periodically to identify elephant flows. In online flow analysis, packet headers or payloads of active flows are analyzed as new flows enter a provider's IP router for a live detection of elephant flows. Finally, in an

end-host assisted approach, software is executed on data-transfer nodes or external nodes that assist in identifying elephant flows and notifying HNTES servers.

For the second task, *circuit provisioning,* HNTES can either request the setup of rate-unlimited MPLS LSPs offline or online, or rate-specified MPLS LSPs online. Typically, if the offline flow analysis approach is used for elephant flow identification, rate-unlimited static LSPs will be configured offline as well, grouping all elephant flows between the same pair of routers on to the same LSP. If online flow analysis is executed, an online circuit provisioning step may be required if a circuit does not already exist between ingress-egress routers corresponding to this flow. If an end-host assisted approach is used for elephant flow identification, when HNTES learns about the imminent start of an elephant flow from an end host, it can ask for a rate-specified or rate-unlimited circuit based on the specifications received from the end host. A more detailed explanation on the usefulness of rate-unlimited LSPs is provided in Section 6.2.

The third task, *PBR configuration*, can be done offline (a priori) or online (on flow arrival).

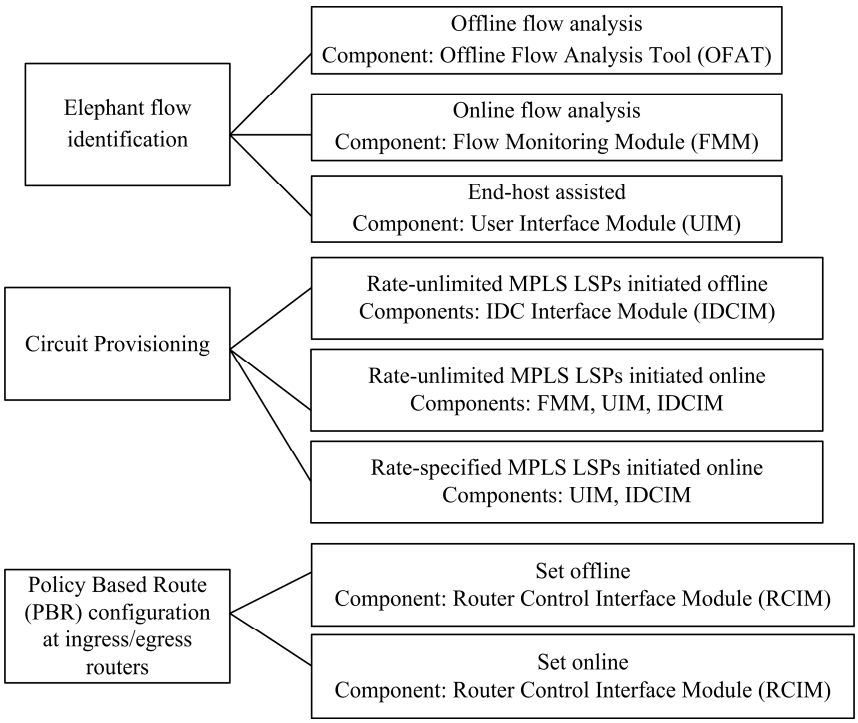

Figure 3: Tasks executed by HNTES

5.3 HNTES architecture

The six modules of the HNTES architecture are shown in Figure 4. The external entities to which HNTES connects are the IP Routers shown at the bottom of the figure, and the OSCARS IDC shown on the right-hand side of the figure. The Router Control Interface Module (RCIM) connects to the control ports of the routers as it only sends configuration commands for setting policy based routes. The Flow Monitoring Module connects to the data ports of the routers because packets arriving on other ports are mirrored to the FMM ports. The tasks performed by each module are shown in Figure 3.

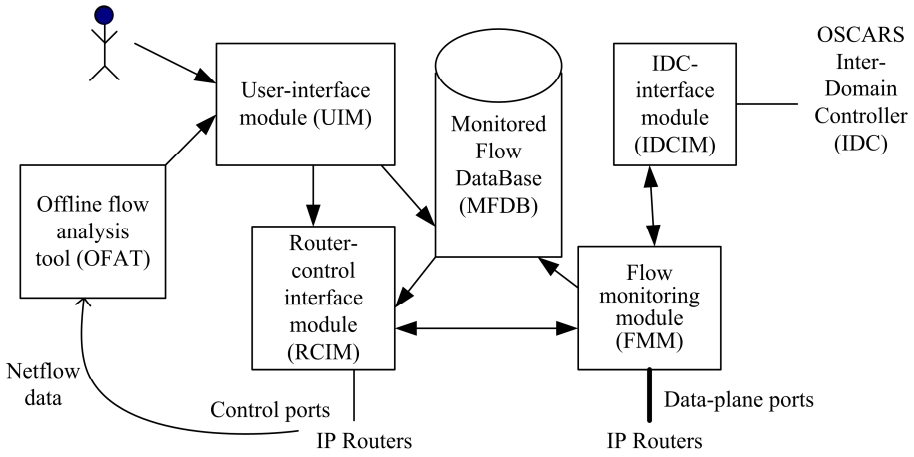

Figure 4: Hybrid Network Traffic Engineering System (HNTES) architecture

6 Currently funded work

The University of Virginia is currently funded by a DOE ASCR grant to design and prototype a Hybrid Network Traffic Engineering System (HNTES). All developed software and documents are available through the project web site [27].

This section explains what has been accomplished to date on the currently funded project (which was started in Sept. 2009), and what will be completed in the time left in this project (which ends in Aug. 2012).

6.1 Year 1 (Sept. 2009 - Aug. 2010)

Our focus was on **dynamic circuits**, and hence on online flow detection, live circuit setup, and live flow redirection ("live" implies that actions are taken after the flow is detected). Also, the heavy-hitters were chosen to be **long-duration flows**.

- **Netflow data analysis:**
  - Since OSCARS circuit setup delay was on the order of 5 minutes (it has been reduced recently to 1 minute), we decided that long-duration flows were the only ones that could use circuits because of our focus on dynamic circuits. Therefore, we developed algorithms and implemented statistical data analysis programs to analyze Netflow data to find long-duration flows. Internet2 Netflow data was analyzed because ESnet Netflow data was unavailable to us.
- **HNTES 1.0 design:**
  - A high-level hybrid network architecture document [28] and a detailed Hybrid Network Traffic Engineering Software (HNTES[b]) design document [27], describing the components shown in Figure 4 were completed.

[b] Recently, we replaced the last word "Software" with "System" in the acronym HNTES.

- **HNTES 1.0 software development:**

  Different modules of the HNTES architecture were prototyped. This design called for "live" flow detection, "live" circuit setup, and "live" flow redirection.

  - Step 1: Identifiers of long-duration flows found through Netflow analysis, executed by the Offline Flow Analysis Tool (OFAT) module of the HNTES software (see Figure 4), are stored in an SQL Monitored Flow Data Base (MFDB), and IP routers are configured by the Router Control Interface Module (RCIM) to port mirror packets from all of these flows to the Flow Monitoring Module (FMM) (see Figure 4).
  - Step 2: When a packet from one of these monitored flows arrives at an IP router, it is port mirrored by the router to the HNTES server, and is thus received by the FMM (this is "live" online detection). The FMM then sends a message to the IDC Interface Module (IDCIM) to initiate circuit setup (see Figure 4). The latter communicates with an OSCARS IDC (this is "live" circuit setup).
  - Step 3: After successful circuit setup, the IDC configures a policy based route (PBR) at the IP router to redirect packets for the live-detected flow to the newly established circuit (this is "live" flow redirection).

- **HNTES 1.0 testing on DOE Advanced Networking Initiative (ANI) testbed:**
  - We executed experiments on the DOE ANI testbed [29] to test modules of the HNTES 1.0 software with support from ESnet.
  - We demonstrated the code with video recordings (posted on the project web site [27]) in an Oct. 2010 review with the DOE program manager. An interesting set of experiments tested the possibility of out-of-sequence packets when flows are redirected from IP-routed paths to circuits. When rate limiting is applied to the IP-routed path on the testbed to lower the link rate to below that of circuit emulated path, then packets did arrive out of sequence, causing TCP throughput to drop. Three duplicate acknowledgments will cause the TCP sender to flag this event as a lost segment and hence half the congestion window size. This points to a need for careful selection of circuit rates before redirection of packets on live flows.

6.2 Completed part of Year 2 (Sept. 2010 - present)

We changed focus from long-duration flows to **large-sized (elephant) flows**, and from dynamic circuits to **static circuits**, the reasons for which are explained below.

- **Identified issues with the HNTES 1.0 solution and adopted a shift in thinking:**
  - Heavy-hitter science flows on ESnet and DOE laboratory networks are primarily of the large-sized variety. Furthermore, simulation studies showed that long-duration flows with low-to-moderate rate circuits (which are sufficient for low-latency applications such as remote visualization [30]) have smaller effects on general-purpose flows than large-sized flows.
  - Experiments (see Section 8.1) conducted between hosts located at ESnet Points of Presence (PoPs) showed that 100 MB transfers take only 2 seconds each, and 1 GB transfers complete in 10 seconds each. While the threshold for a file to qualify as

"large-sized" could be considerably bigger, nevertheless, the transfer delays could be smaller than circuit setup, which takes on the order of minutes. Therefore, the solution considers the use of static circuits, i.e., pre-provisioned circuits.

- The Year 1 focus was on creating one circuit per science flow (fine granularity) since the goal was to provide science flows rate-guaranteed circuits. In year 2, our focus is on isolating elephant flows from general-purpose flows by moving them off IP-routed networks to circuits for reasons provided in Section 2. For this goal, individual per-flow circuits are not required. Instead, all elephant flows between the same ingress-egress ESnet routers can be grouped together onto one Label Switched Path (LSP). This is a coarse grained approach for using circuits for science flows. Such an approach is feasible with static circuits.
- Flow Monitoring Module (FMM) in HNTES 1.0 is impractical because it needs to process headers from all mirrored data packets for heavy hitter flows in real time, and as link speeds increase, this becomes infeasible without a high-performance computing platform. Requiring the deployment of such platforms to serve one or more ESnet routers could be cost prohibitive. Therefore, online (live) flow detection solutions that require the examination of even just the headers of data-plane packets are not desirable.

- **Innovative approach to use rate-unlimited static LSPs:**
  - If static circuits used between ingress-egress router pairs of ESnet or any other network are made to carry aggregate flows for science flow isolation reasons, the question is whether these circuits should be rate limited. For example, if each ESnet router is to be connected to its other routers via static circuits, in today's structure, each 10 Gbps link has to be divided by some large number, such as 50 if there are 41 other sites and at least 9 border routers that peer with other providers, making each circuit rate quite low. The science flows will suffer from low throughput if such as approach is adopted. On the other hand, if all LSPs could be allowed to enjoy full link bandwidth (i.e., there is no rate limiting), then science flows would receive the best throughput possible making it counter-productive to isolate each science flow in its own (necessarily lower-rate) LSP. This solution not only moves science flows off the IP-routed links, which means they no longer negatively impact the performance of general-purpose flows, it simultaneously allows science flows to receive the best throughput possible without micro-management of circuit network bandwidth. What is sacrificed is throughput variance, especially if the science-flow loads are moderate to high. This solution is feasible with the MPLS technology through the use of **rate-unlimited static LSPs**.
- **Preliminary HNTES 2.0 design: Combines size dimension with static LSPs**
  - The offline version of all three tasks shown in Figure 3 will be executed, i.e., (i) offline flow analysis to determine large-sized flows, (ii) a periodic (e.g., nightly) run to determine if any new static LSPs are required (if all elephant flows recorded on a day are already in the database then no circuits will be required), and if a new circuit is required, the IDCIM software would contact the IDC to set up circuits, and (iii) a periodic (e.g., nightly) run of the RCIM to add policy based routes to redirect

any newly identified flows to appropriate LSPs. Therefore, none of the three steps: flow identification, circuit setup, and PBR configuration are "live." Further details are provided in Section 6.4.

- **Design algorithms for flow size estimation**
  - Flow size estimation from Netflow data: Section 8.1 describes our experiments to determine whether the size of a flow can be estimated accurately from Netflow data, which consists of sampled packet headers (e.g., ESnet routers execute 1-in-1000 packet sampling).
- **Study of the question "why move elephant flows off the IP-routed network"**
  - Experimental studies across ESnet to see the effect of elephant flows on link loads (passive SNMP measurements): Section 8.2 describes our findings.
  - Simulations: Section 8.3 describes our ongoing simulation studies to test the hypothesis that under moderate-to-high loads, elephant flows consume an unfair share of link bandwidth thus effecting mice flows. Additional simulations will be run to study interactions between elephant flows and delay-sensitive flows.
  - PerfSONAR active measurements analysis: Section 8.4 describes our analysis of PerfSONAR One-Way Ping (OWAMP) [31] data to determine if there are delay surges that can be attributed to router buffer buildups, and to further check if these delays are attributable to elephant flows by correlating with Netflow data.

6.3 Remaining period (present - Aug. 2012):

- **Data analysis of ESnet Netflow data**
  - Section 6.4 describes how a nightly offline analysis will be executed on each day's Netflow data from a sample set of ESnet routers. The goal is to determine if offline flow detection is sufficient for finding elephant flows in ESnet traffic.
- **Complete ongoing work**
  - The simulation studies and PerfSONAR OWAMP analysis as described in Sections 8.3 and 8.4 will be completed.
- **Refine HNTES 2.0 design and prototype**
  - More details about the HNTES 2.0 design are provided in Section 6.4. Input to the design process will be findings from the data analysis.
  - Implement a new version of the OFAT, RCIM, IDCIM, and MFDB modules of HNTES given the shift to elephant flows and static circuits.
- **HNTES 2.0 testing on the ANI 100G testbed**
  - The concept of using rate-unlimited MPLS LSPs and the impact of this type of circuits on science flow throughput will be studied. Experiments will be planned to create multiple simultaneous elephant flows.
  - The execution time for various tasks in HNTES 2.0 will be measured.

6.4 Preliminary HNTES 2.0 design

For **elephant flow identification**, a periodic (e.g., nightly) offline analysis of Netflow data will be executed by the OFAT module. The *algorithm* is as follows. The first problem arises from an engineering aspect of Netflow. Typically, Netflow is configured on IP routers to have short "active timeout intervals," such as 60 seconds, which means flow information is exported every 60 seconds. To determine the total size of a flow, as identified by its 5-tuple (see Section 2), sizes (in bytes) from the multiple reports corresponding to that flow need to be summed. However, due to sampling (e.g., Netflow in ESnet routers is configured to sample 1-in-1000 packets), one cannot determine whether reports for the same flow identifier correspond to the same flow or not, because a flow with a particular identifier may have ended and a new flow initiated with the same identifier between the generation times of flow reports. Also, as per experiments described in Section 8.1, the accuracy of flow-size estimation from such flow reports decreases with decreasing file sizes.

Therefore, we redefine a flow to be identified by *just the source/destination IP address pair*, and add the reported number of bytes from all flow records corresponding to each such pair. In other words, we abandon trying to find the exact flow size for a 5-tuple identified flow, and focus instead of finding the total number of bytes exchanged between source/destination host pairs. The reason for this design choice is that while port numbers are ephemeral in most files transfers, typically particular hosts (single or cluster nodes) are engaged in file transfers from the large dataset serving data-transfer nodes. By adding the byte sizes from all reports corresponding to a flow, as defined by its source/destination IP address pair, we could be adding in bytes contributed by control-plane packets. But the goal here is not to estimate the total size of files transferred, but rather just to *find the flows with some large number of bytes* transferred on some periodic basis. For example, this summation process could be executed every night on a day's worth of Netflow collected files (typically, the Netflow collector receives files every 5 minutes, and therefore there are 288 files per day per router) to determine the identifiers (source/destination IP address pairs) of all flows that exceed 1 GB.

Table 1: Elephant flow database; as an example, period is one day and the sliding window is 30 days.

Row number	Source IP address	Destination IP address	Is the source a data door?	Is the destination a data door?	Day 1	Day 2	....	Day 30
					(total transfer size; if one day the total transfer size between this node pair is < 1GB, list 0)			
1			0 or 1	0 or 1				
2								

The *next step* is to store information about these flows whose sum total of bytes over a defined period, e.g., a day, is larger than a prespecified size threshold, e.g., 1 GB. A multi-day counter is used as sliding window, which means a new entry overwrites the most dated entry. For a new flow, the entry is written for that particular day, but "NA (not available)" is recorded for all other days. If the amount of data transferred for an existing flow in Table 1

does not exceed the prespecified size threshold, e.g. 1 GB, then that value is entered as 0. The rationale for the "NA" marking is that if a new scientist joins a project and starts downloading data on a certain day, then that flow should not be discounted in the next step of elephant flow identification in which a persistency measure in used.

Two additional columns are stored in this table to track whether or not the source and destination IP addresses are known "data doors." The term *data door* is used to represent data transfer nodes that have been explicitly tasked for data transfers from the major computing facilities run by the DOE laboratories and research universities (e.g., dtn02.nersc.gov). Correlating the elephant flow information with our database of data door IP addresses and storing this information, as in Table 1, provides us an understanding of the source/destination of the elephant flows.

In this *next step*, at the end of every period (e.g., every night), the data in Table 1 is sorted by a weighted sum of the percentage of days (out of the days for which data is available for each flow) in which the size is non-zero (persistency measure), and the average per-day size, with the averaging also taken over the number of days for which data is available for the flow (size measure). A threshold is set for this weighted sum. Flows that do not meet this threshold are discarded from the table. This step is required to keep the table from growing too large. A higher weight is assigned for the persistency measure if the cost of online flow detection, live circuit setup and flow direction is relatively low. But if this cost is high, then the size measure should be given a higher weight.

Circuit provisioning: For any newly added flows in the table, using a combination of interface indices in Netflow reports, traceroute, and PerfSONAR traceroute, the IDCIM module of HNTES determines the ingress and egress IP routers within the provider's network corresponding to the source and destination IP addresses of each flow. This information is used to populate the database in Table 2. A lookup of the database yields whether or not a circuit already exists between the ingress-egress router pair. If it does, the IDCIM sends a message to the RCIM for the PBR configuration step. If not, it requests a new circuit from the IDC.

Table 2: Circuits and Monitored flow database

Row number	Source IP address	Destination IP address	Ingress IP router	Egress IP router	Circuit exists?
1					
2					

Policy Based Route (PBR) configuration: This step is executed by the RCIM, which is notified by the IDCIM, if any new flows require the addition of policy based routes.

The above is a preliminary HNTES 2.0 design. As shown in Section 6.3, the HNTES 2.0 design will be refined based on the data analysis step. Parameters, such as periodicity of offline execution of the HNTES modules, number of columns in the database shown in Table 1, and the size threshold, will be defined based on the data analysis.

Note that the FMM is not involved in HNTES 2.0 as there is no online elephant flow detection. This is part of the work proposed for this application (see Section 7).

7 Proposed work

This section describes the new work being proposed in this response to the current DOE solicitation [1], which will extend and build on the currently funded work on HNTES. It includes the following items:

- **Design of new online flow detection algorithms**
- **HNTES 3.0 design, prototyping and testing (superset of HNTES 2.0)**
- **Design of end-host assisted flow identification mechanisms**
- **HNTES 4.0 design, prototyping and testing (superset of HNTES 3.0)**
- **HNTES in an integrated network: Optimal algorithms for load balancing science flows and sizing LSPs**
- **Other types of heavy-hitter flows**

7.1 Design of new online flow detection algorithms

It is our hypothesis that when the HNTES 2.0 offline flow analysis tool is executed on a periodic, say daily, basis using the elephant flow identification algorithm described in Section 6.4, there will be some non-zero percentage of new elephant flows identified on some days. Two histograms can be plotted: (i) percentage of new elephant flows identified each day, and (ii) persistency measure: percentage of days in which a flow appeared in the table (which means its flow size was larger than the size threshold) since it first appeared. If the first histogram is skewed left (long tail on the left side), then HNTES requires an online flow detection module. But if the second histogram is skewed left, then that argues for using primarily an offline flow detection scheme.

While this data will be analyzed for ESnet, a general HNTES solution that is useful for DOE laboratory networks as well as other provider networks should include an online elephant flow detection module. Algorithms are required for this functionality. *Three* possibilities are currently identified. The first is a *packet-header based* scheme, the second, a *payload based scheme* (see Section 4), and the third is based on *Netflow-data analysis*.

In the *packet-header based scheme*, TCP SYN segments destined to well-known data transfer nodes will be port-mirrored from IP routers to the Flow Monitoring Module (FMM) of a HNTES server (see Figure 4). Most routers support port-mirroring specifications that allow only 0-length packets to be sent to the mirror site. This avoids the problem noted in Section 6.2 of the FMM having to keep pace with high-rate data-plane packets. These TCP SYN segment carrying IP packet headers will be parsed to extract the source address allowing for an identification of the source-destination IP addresses of a potential elephant flow. Since TCP data connections need to be established prior to data transfer, using these TCP SYN segments is one way of detecting potential elephant flows online. The implication of some of these flows being mice flows is addressed in Section 7.2.

In the *payload based scheme*, the packets on control connections of common file transfer applications will be mirrored to the FMM. Control ports are often well-known, such as 2811 for GridFTP. While other control port numbers have been detected in use by GridFTP servers in Netflow data, 2811 is a commonly used port number.

We use GridFTP as an example here, but plan to develop deep packet inspection algorithms for the control packets of the other popular file transfer applications, such as BBFTP. This scheme is possible only for file transfer applications in which the control messages are not encrypted. The GridFTP architecture consists of a client PI (protocol interpreter), one server PI at each end, and one or more DTPs (Data Transfer Processes), one per cluster node. Transfers can occur in first-party mode or third-party mode. In first-party mode, the host on which the client PI is being run will typically also be running the DTP. If the server PI and server DTP are also run on the same host, the source and destination IP addresses for the data connection will be the same as those of the control packets, which can be extracted from the mirrored control packet headers. However, for third-party transfers, the data connection IP addresses need to be extracted from commands such as PASV/SPAS and PORT/SPOR. These commands also carry port numbers. But since the latter are ephemeral, these could change from one transfer to another. Therefore, the PBR configurations should be made with just the source/destination IP addresses.

Experiments will be conducted on the ANI testbed using tcpdump to collect control messages for GridFTP, BBFTP, and other applications, to develop specific deep packet inspection algorithms for online flow detection using payload based classification.

Finally, "real-time" analysis of Netflow data is a possibility. Since ESnet routers send flow reports only every 5 minutes, this will incur some delay, but it may nevertheless be useful to leverage this information without having to wait for the over-night runs.

7.2 Design, prototyping and testing of HNTES 3.0 (superset of HNTES 2.0)

HNTES 3.0 adds the Flow Monitoring Module (FMM) to HNTES 2.0 (hence it is a superset). This FMM is considerably different from that in HNTES 1.0 and hence requires a complete redesign. HNTES 2.0 does not include an FMM as noted in Section 6.4. The algorithms described, such as deep packet inspection of control messages, and extraction of addresses from TCP SYN segment carrying IP packet headers, will be coded. As the FMM performs live (online) flow detection, the remaining HNTES modules need to act quickly to set up circuits if required, and redirect flows through PBR configurations.

For the circuit provisioning step, as described in Section 6.4, the functionality of determining the ingress-egress router pairs corresponding to each flow is executed by the IDCIM. This requires fast communication between the FMM and IDCIM. Once the latter determines the ingress-egress router pairs, it needs to look up the Circuits and Monitored Flow Data Base (Table 2) to determine if a circuit already exists. If it does, then the IDCIM has to notify the RCIM (see Figure 4) to set an online PBR in the router for this new flow. If a circuit does not exist, the IDCIM needs to request the IDC to set up a rateless static MPLS LSP (this may require new features in the IDC). The use of rateless static LSPs removes the need for an automatic estimation of circuit rate and circuit duration, both of which are difficult to estimate for a new flow, which is not in the elephant flow database (Table 1). The IDC itself can be asked to execute the PBR configuration action after setting up the circuit (it should be provided the flow identifiers with the circuit request).

This solution combines the "online" solution for elephant flow identification and PBR configuration tasks of Figure 3, and uses either the offline or online rateless static MPLS LSP circuit provisioning solutions.

Performance optimization: This design requires fast circuit provisioning and PBR configuration. Currently both take on the order of minutes, which means the online detected flow is likely to complete before it can be redirected. It is still useful to put this redirection in place for two reasons: (i) some of the detected flows may indeed be long-lasting elephant flows, and (ii) the probability of other data transfers being initiated between the same source-destination pair is likely to be high. Since circuits are not being micro-managed but instead multiple flows are directed to the same circuit, redirection of some mice flows (just those destined to the data transfer nodes) is not likely to cause significant problems. Nevertheless, we plan to investigate the use of OpenFlow switches, and other techniques for decreasing circuit provisioning and PBR configuration times.

7.3 Design of end-host assisted flow identification mechanisms

The concept here is to have end hosts (e.g., the data transfer nodes themselves or external hosts located in the campus networks) assist in elephant flow identification. We use the name *end-host flow detection module (EFDM)* for the software program that performs this functionality. If the EFDM is executed on the same data transfer nodes as the file transfer applications, it can use the pcap library to capture TCP segments sent out of or received on the network interface cards and analyze control messages, and or TCP SYN segments to identify elephant flows. For example, if the control messages are unencrypted, the EFDM can extract the names of the files being requested and determine the sizes of the files. It can thus identify exactly which transfers are elephant flows. If the EFDM is run on an external host, the same port mirroring concept used for the FMM can be used here. Another approach is to use mechanisms such as those developed by Phoebus [20], which capture OS socket calls and/or use iptables. These different approaches will be evaluated.

Using traceroute and/or PerfSONAR [32] traceroute, combined with the PerfSONAR lookup service with which HNTES servers operated by providers can be registered, the EFDM can notify any en route HNTES servers about the imminent start of an elephant flow. This is also an "online" elephant flow identification mechanism and corresponds to the third option shown in Figure 3. The User Interface Module (UIM) (Figure 4) then interfaces with the IDCIM and RCIM to provision a new circuit and add a new PBR, if needed.

7.4 Design, prototyping and testing of HNTES 4.0 (superset of HNTES 3.0)

HNTES 4.0 will include the above-described option of end-host assisted flow identification. HNTES 4.0 adds new functionality to its UIM module (hence it is a superset of HNTES 3.0), and adds an EFDM to be executed at data transfer nodes or on separate end hosts in campus networks. The modules will be tested on the ANI 100G testbed.

Whether such a solution is feasible depends upon the details of the commonly used file transfer applications, as well as performance requirements. A large-scale cluster of data transfer nodes will have a high rate of file transfer requests. The throughput performance of the EFDM will need to be estimated from per-task execution times using a model. One advantage of this solution over the FMM based online flow detection is that at least for some file transfer applications, the EFDM can know the file size and therefore send requests to HNTES servers only for elephant flows. Also the UIM could determine, through other means, an appropriate circuit rate, estimate circuit duration (as it knows the

cumulative size of the files to be transferred), and make requests for rate-specified circuits as shown in the third option for the circuit provisioning task in Figure 3.

7.5 HNTES in an integrated network

Section 2 described two options for a hybrid network: (i) separate networks to support the IP-routed and services, as is deployed in today's ESnet4, and (ii) an integrated network as is one of the options being considered for the next-generation ESnet with 100 Gbps links. The concept of using rate-unlimited static LSPs, interconnecting ingress and egress routers of ESnet or any other provider network was introduced in Section 6.2. Given the short durations of elephant flows relative to circuit provisioning delays, static circuits are required. The advantage of rate-unlimited LSPs in which all elephant flows are batched together (i.e., high throughput for science flows unfettered by rate limits) was also described in Section 6.2. This solution is easy to envision in the separate network scenario. For example, in Figure 2, a rate-unlimited static LSP would be provisioned between R3 and R2 traversing MPLS LSRs S3 and S2. This path is physically distinct from the IP-routed path between R3 and R2.

The research challenge addressed here is to determine mechanisms for realizing the above solution of rate-unlimited static LSPs in an integrated network. While rate-specified circuits may be requested by the end-host module as described in Section 7.4 for HNTES 4.0, here we consider the independent network provider solutions of HNTES 2.0 and 3.0 in which HNTES through offline or online flow identification schemes decides to redirect science flows off the IP-routed paths.

Two solutions are identified for supporting HNTES 3.0 (which includes the offline and online provider-only driven approaches) in an integrated network: (i) two logically separate IP-routed networks; (ii) one logical IP-routed network and one logical MPLS network. The second uses an innovative idea to obtain traffic matrices that are required for sizing the two logical network links. *Other possible solutions* will also be investigated.

In the *first solution*, a network provider would create two logical IP-routed networks by configuring two single-link MPLS LSPs on each link of the whole network (e.g., on each link of the ESnet topology). One IP-routed network would be used to carry general-purpose flows and the other to carry science flows. The first set of LSPs for the general-purpose flows would be rate limited to some value, e.g., 20%, of link bandwidth, while the second set of LSPs for science flows would be rate limited to the remaining link bandwidth (e.g., 80%). Effectively this creates two logical networks, but both operate as IP-routed networks. Packets sent on the single-link science LSPs will need to be reassembled and examined at the IP layer at each router, and the router will determine the next hop by consulting its IP routing tables. The single-link MPLS LSPs are virtual interfaces. In this solution, science flows are isolated to their own network, but there is no usage of multi-link circuits.

The *second solution* requires the provider to similarly divide each physical link into two logical links using single-link MPLS LSPs. But it further requires the creation of a second MPLS layer on top of the science single-link LSPs using the label stacking feature of MPLS. This second layer would be used to create the rate-unlimited static LSPs between ingress-egress router pairs. Science flows would be switched at the MPLS layer at transit nodes

rather than at the IP layer as in the first solution. This solution effectively emulates the solution proposed for HNTES 2.0 in the separate network scenario.

Comparison of the solutions: In the first solution, there is no effective way to achieve load balancing for the science flows. For example, if flows from Sunnyvale to New York are routed to Denver, they would compete with flows from Sunnyvale to Denver, Sunnyvale to Kansas City, Sunnyvale to Chicago, etc. But with the second solution, a traffic engineering (TE) system can spread the load to a larger number of links. A *key concept*, obtained from [33], is to use rate-unlimited MPLS LSPs fully connected in a logical mesh to determine traffic matrices from SNMP link usage measurements. These SNMP measurements can be obtained for the logical links, i.e., the MPLS LSPs. Obtaining a traffic matrix from an IP-routed network without such a mesh of rate-unlimited LSPs is difficult [34]. But with the logical mesh of LSPs, the TE system can easily construct a provider-wide traffic matrix for the science flows. It can then run an optimization algorithm to determine optimal routing of these MPLS LSPs so that the load is spread out among the links of the network. Effectively this means, that at a given utilization, science flows will enjoy higher throughput than in the first solution because the TE system can periodically adjust the routing of the top-layer LSPs to ensure that science flows are directed to links with lower loads whenever possible. Simulations will be executed to compare these two solutions, and both solutions will be tested on the ANI 100G testbed.

The second solution also allows for an easier solution to the network management issue of how to size the single-link MPLS LSPs (i.e., how to choose the example 20% value for IP-routed networks). This problem will also be addressed in this work.

7.6 Other types of heavy-hitter flows

The offline Netflow analysis will be extended to find other types of heavy-hitter flows such as high-rate, long-duration or excessively bursty flows. Simulations will be undertaken to decide which of those, if any, should be redirected off the IP-routed networks.

Also further explanations will be sought for persistent large-sized flows that cannot be attributed to specific data doors (as in Table 1). Such flows could be due to general-purpose transfers such as Linux update distributions, or other peer-to-peer applications that are not scientific in nature. These findings should be useful to ESnet.

Finally, the use of complete source/destination IP address pairs (/32) to identify flows will be generalized to variable-length subnets. The impact of routing all flows between subnets on to circuits will be studied.

8 Preliminary results

8.1 Flow size estimation experiments

We executed an experimental study across ESnet to test the conditions under which the hypothesis that "a flow size from sampled Netflow records can be multiplied by 1000 (since the ESnet Netflow sampling rate is 1 in 1000 packets) to estimate actual flow size" holds. We ran GridFTP transfers with files of known sizes, and collected Netflow data to test the hypothesis. For large files, we expect that the hypothesis will be true, and false for

small files. The hosts *anl-diskpt1* and *lbl-diskpt1* located at ANL and LBL, respectively, were used to run the GridFTP server and client. There are eight ESnet IP routers in between these hosts, with 10 Gb/s links between the routers.

A workflow script is created to automate the steps involved in this experiment. First a firewall counter is read at an en route IP router. Next the GridFTP transfer is initiated. When the transfer completes, the firewall counter is re-read to verify the transferred file size in bytes. From the GridFTP logs stored at the server, the data ports used for the TCP connections are obtained. Netflow data is collected from two of the routers. All flow records corresponding to the GridFTP transfer are filtered out of the Netflow records using the five-tuple identifier, and the size of the transfer is estimated using the 1000 factor multiplier on the total size reported by the flow records for the data connection. The *size-accuracy ratio* is defined to be the ratio of the Netflow estimated size and actual file size.

For each file size (100MB, 1GB, and 10 GB), multiple runs were executed since the packet sampling at the router makes the size-accuracy ratio a random variable. The results obtained are shown in Table 3. For all three file sizes, the sample mean shows a size-accuracy ratio close to 1, and more interestingly, the standard deviation is smaller for larger files. As these experiments were run across ESnet, traffic loads could be quite different when each of the runs was executed. We are planning to correlate this data with SNMP link loads at the times of the runs to study the impact of load on Netflow size estimation accuracy.

Table 3: Size-accuracy ratio for files for different sizes (sample size: 50)

	Netflow records obtained from Chicago ESnet router		Netflow records obtained from Sunnyvale ESnet router	
	Mean	Standard deviation	Mean	Standard deviation
100 MB	0.949	0.2780	1.0812	0.3073
1 GB	0.996	0.1708	1.032	0.1653
10 GB	0.990	0.0368	0.999	0.0252

8.2 Experimental study of the effect of elephant flow on link loads

To check whether elephant flows have an effect on link loads, SNMP link usage measurements (which averages bits sent/received over 30 sec periods) are obtained for the time interval in which five 10 GB GridFTP transfers were executed in the experiments described in Section 8.1. Figure 5 shows the SNMP link load plots while the GridFTP transfers were executed. The y-axis is the link load from SNMP measurements. The first marked horizontal line is 2.5 Gbps. Link capacity is 10 Gbps. The x-axis indicates time vertical lines separated by 1 minute. The solid lines show SNMP link loads on two router interfaces (on the Chicago and Sunnyvale ESnet routers) including the five experimental 10 GB GridFTP transfers, and the dashed lines represent the rest of the traffic load. The increases in link loads caused by the five experimental 10 GB transfers are clearly evident.

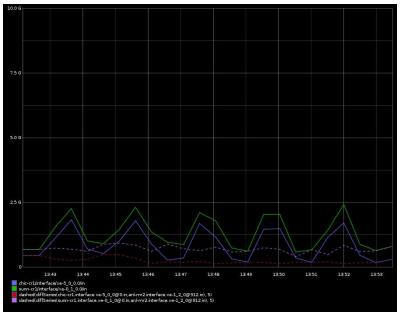

Figure 5

8.3 Simulations

The goal is to use simulations to study the effect described in Section 2, whereby because of TCP's congestion control algorithm, elephant flows last long enough to start consuming larger portions of the link bandwidth while mice flows terminate before their TCP sending window size can grow to the bandwidth-delay product size, especially for high-speed, long-distance transfers. On the other hand, the arrival rate of mice flows is significantly higher than that of elephant flows. This is because file size has been characterized to fit a Pareto distribution [35] (whose density function is shown Figure 6. The *k* value is for the shape parameter (the larger the *k*, the higher the probability of small files). The other parameter, the scale parameter, which is the smallest file size, is chosen based on examples in [36]; with these choices, the mean file size is only 33.8 KB.

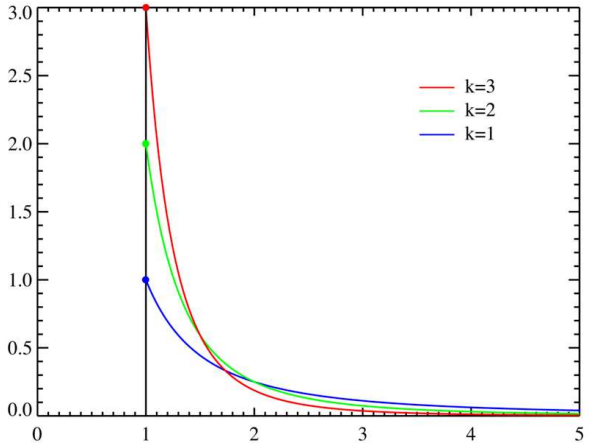

Figure 6: PDF of a Pareto random variable (http://en.wikipedia.org/wiki/File:Pareto_distributionPDF.png)

To test the effect of these two counteracting forces, we simulated an ideal equal bandwidth sharing algorithm in which packets from flows are served in a round-robin fashion (this is not practical in today's routers as packets are not separated out into flows) so that all concurrent flows receive an equal share of link bandwidth. We also simulated an unequal bandwidth-sharing algorithm in which elephant flows (a file size threshold was set to classify flows as mice or elephants) received four times as much bandwidth as mice flows. The factor 4 was chosen arbitrarily. Sensitivity analysis can be carried out for this factor as well as the threshold parameter.

The file-transfer arrival process is assumed to be Poisson, as validated by earlier work [37]. The fairness metric used is as follows [38]:

$$f(x) = \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n\sum_{i=1}^{n} x_i^2}, \text{ where } x_i \geq 0.$$

For our purposes, x_i is mean throughput (rate) and Inter-Quartile Range (IQR) of the mean rate. The mean rate and IQR of mean rate are computed across the values obtained in 1000 simulation runs. Files are grouped into file size bins (in KB), < 12, 12-15, 15-20, 20-30, 30-73 and > 73. These sizes are chosen to guarantee that there are at least 1500 files in each bin (from the total 10,000 files simulated on each run). Corresponding to these bins, there are 6 classes (i.e., n=6). The equal and unequal bandwidth sharing algorithms are applied to files whose sizes and inter-arrival times are random numbers generated with a statistical R [39] program to fit the assumed distributions.

Table 4: Fairness metric under equal (1:1) and unequal (4:1) bandwidth-sharing algorithms

	Mean rate under different loads				Mean rate IQR under different loads			
	0.81	0.67	0.54	0.27	0.81	0.67	0.54	0.27
1:1	0.99750	0.99808	0.99876	0.99966	0.99651	0.99866	0.99974	0.99680
4:1	0.98246	0.99169	0.99596	0.99938	0.99706	0.99326	0.98984	0.98594

The mean rate itself is higher for smaller files even in the unequal bandwidth sharing case because of the larger number of mice flows. This effect is even more for larger values of *k,* the shape parameter.

The system is less fair at higher loads when the fairness metric uses mean rate. This is because at higher loads, more mice flows will arrive at the link within the lifetime of an elephant flow, causing its transfer duration to keep getting extended over time.

The next steps are to simulate the TCP bandwidth sharing algorithms in ns2 [40], and to add a new class of file arrivals in which the mean file size is large (on the order of 100MB) to capture the much larger scientific datasets. While the Pareto distribution has a long tail, the probability of a 100 MB file with the parameters chosen for the general-purpose file sizes is small, requiring simulations to be run for a long duration before a statistically significant sample of large files can be obtained. Our conclusion is that we require two separate file arrival processes, one that characterizes general-purpose flows and the other that characterizes science flows. This is our planned next step (see Section 6.3).

8.4 PerfSONAR OWAMP analysis

As noted in Section 6.2, to answer the "why move elephant flows off the IP-routed network," we decided to look for surges in one-way delay in PerfSONAR OWAMP measurements. The hypothesis is that if an elephant flow causes router buffer buildups, there will be periods in which the OWAMP delays are higher than the minimum. One-Way Ping (OWAMP) servers are deployed by providers such as ESnet and Internet2. The system clocks at the two ends are synchronized. Internet2 dedicates “latency hosts” to just run OWAMP tests and collect measurements. These measurements are periodically pushed to a measurement archive. Twenty packets are sent per second on average (10 for ipv4, 10 for ipv6) to each of the other OWAMP servers (there are 9 on Internet2). We were able to obtain raw OWAMP measurements (1200 per minute; as opposed to just the maximum and minumum values per minute that are available through PerfSONAR Web sites) for two weeks for all pairs. We defined a "surge point" on a per-minute basis as follows. Compute (i) b, the 10th percentile delay across the 2-week data, and (ii) i, the per-minute 10th percentile delay. Using a factor n, a surge point is defined as one whose delay $i \geq n \times b$. Consecutive surge points are combined into a single "surge."

The results that showed that the surge duration (width), size (height), and frequency of surges vary for different paths and different definitions of the term “surge point” (value of n). For one path between the Los Angeles and Salt Lake City OWAMP servers, using n = 1.5 in our definition of surge point, we found that the maximum surge duration over the two

week period is over 200 minutes, and that the median is 34 minutes. Figure 7 shows the probability density function of surge duration for this path.

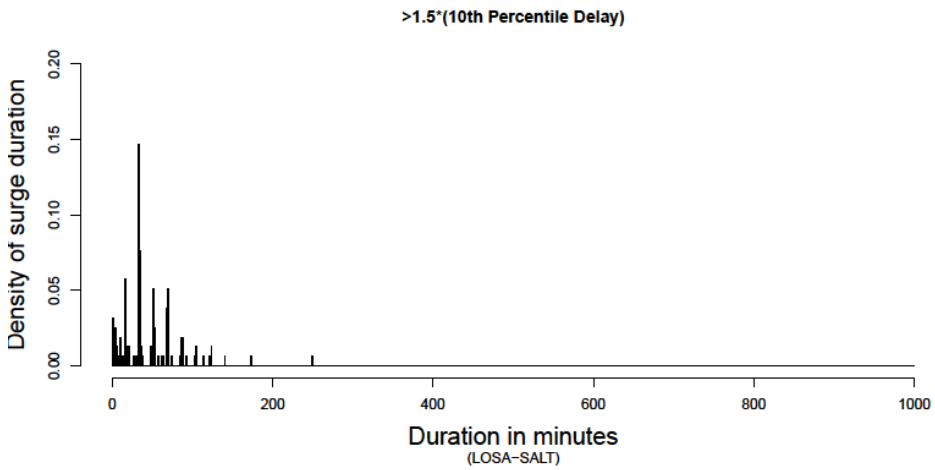

Figure 7: Probability density function of surge duration where the factor n in the definition of surge point is 1.5

The next step is to identify the cause of these OWAMP delay surges. One idea is to correlate these surges with Netflow data that show the time ranges when elephant flows are present (as per the experiments in Section 8.1, since elephant flows have more packets, they are more likely to be sampled by Netflow). If such correlations exist, we can see evidence of the problems caused by elephant flows.

9 Evaluation plan

9.1 Tasks and milestones on a timeline

In the first column, V= U. of Virginia; E = ESnet.

Team	Milestone	Month
E/V	Run experiments on ESnet or ANI testbed to capture control-connection packets for file transfer applications (Section 7.1)	4
E	Run the Netflow data analysis scripts to find histograms (Section 7.1)	6
V	Design of new online flow detection algorithms (Section 7.1)	6
E	Review/improve design of online flow detection algorithms (Section 7.1)	8
V	HNTES 3.0 prototyping and testing (Section 7.2)	12
V/E	Run experiments on ANI testbed to capture tcpdump files for file transfer applications at the data transfer nodes (Section 7.3)	15
E	Run scripts provided by V for Netflow analysis to identify other types of heavy hitter flows (Section 7.6)	18
V	Design of end-host assisted flow identification mechanisms (Section 7.3)	18
E	Review/improve design of end-host assisted flow identification mechanisms (Section 7.3)	21

V	HNTES 4.0 prototyping and testing (Section 7.4)	24
E	Continue Netflow data analysis scripts execution (Section 7.1)	24
E	Run scripts provided by V for Netflow analysis to identify other types of heavy hitter flows (Section 7.6)	30
V	HNTES in an integrated network: Optimal algorithms for load balancing science flows and sizing LSPs (Section 7.5)	30
E	Review design for integrated network solution (Section 7.5)	33
V	Complete all documentation and post final software modules on Web site	36

9.2 Management plan

The University of Virginia (UVA) PI and the ESnet PI have already been collaborating for a year on problems related to the current HNTES project (ESnet is not funded through the HNTES grant, but as the problems being addressed by this project are of interest to ESnet, this collaboration has been ongoing). We meet once a week via a skype video call. We are currently co-authoring a paper on our findings. The work reported in Sections 8.1 and 8.2) are a result of this collaboration.

This meeting format will continue to be used in this proposed project. In addition, the UVA graduate student will also participate in these weekly video calls. As in the tasks and milestones listed in Section 9.1, ESnet will review all algorithms and system designs and provide UVA with feedback based on real-world constraints. We thus expect this collaboration to result in prototypes that could be deployed if appropriate in a follow-up advanced deployment project.

UVa has a Web based Collaboration facility that is available to faculty and students for teaching and research purposes. With Drop boxes, Resources, Forums, Calendar, and various other tools, it is already used extensively by the PIs for the current research project. It allows for external collaborators to also be supported via non-UVA email addresses. The ESnet PI can be listed as a project member, and can post material, join discussions, etc. This will be the main tool to collate and share information between UVA and ESnet.

9.3 Metrics

Our proposed measures of success include demonstrations of HNTES 2.0, HNTES 3.0 and HNTES 4.0 on the ANI 100 G testbed, research publications of our new hybrid networking algorithms, Netflow data analysis, experiments and simulation studies.